JPB REVIEW
INTERNATIONAL JOURNAL OF PROFESSIONAL BUSINESS REVIEW
ISSN: 2525-3654
OPEN ACCESS

PHISH OR LEGIT? QUANTIFYING WEB-BASED THREAT SIGNALS THROUGH PREDICTIVE ANALYTICS AND FEATURE ATTRIBUTION

Daniel Duah[A], Bismark Kofi Owusu Sarfo[B], Collins Boakye[C], Dina Duku[D]

ARTICLE INFO

Article history:

Received: Feb, 14th 2025

Accepted: Apr, 14th 2025

Keywords:

Phishing;
Cybersecurity;
Structural;
Behavioral;
Domain-Based.

ABSTRACT

Objective: This study investigates web-based threat signals using predictive analytics and feature attribution to determine whether a webpage is phishing or legitimate.

Theoretical Framework: The research is grounded in Protection Motivation Theory (PMT), which offers a behavioral lens to interpret phishing indicators. PMT connects web features to users' cognitive threat and coping appraisals, providing a theoretical rationale for selecting and organizing features.

Method: A logistic regression model, regularized with L1 (Lasso), was developed for its interpretability and ability to handle feature sparsity and convergence issues. Using a dataset of 11,055 labeled websites, the model incorporates three core feature sets: structural (e.g., IP-based URLs, SSL status), behavioral (e.g., redirection, form handler anomalies), and domain metadata (e.g., traffic rank, Google indexing).

Results and Discussion: The model rejects the null hypothesis that website-level features are non-predictive, confirming that structural, behavioral, and metadata-based signals significantly distinguish phishing from legitimate sites. This thematic decomposition supports both the conceptual framework and the empirical model design.

Research Implications: The findings offer actionable insights for cybersecurity professionals, especially those in regulated industries. The model enhances detection capability while maintaining transparency, crucial for compliance and risk management.

Originality/Value: This study contributes to literature by integrating PMT into a predictive modeling framework for phishing detection, an approach that bridges behavioral theory and machine learning. Its originality lies in aligning cognitive appraisal theory with interpretable statistical methods. The results are highly relevant to cybersecurity practice, offering scalable, transparent tools that support real-time decision-making and inform strategic defenses in high-risk sectors.

Doi: https://doi.org/10.26668/businessreview/2025.v10i5.5472

PHISH OU LEGÍTIMO? QUANTIFICAÇÃO DE SINAIS DE AMEAÇAS BASEADAS NA WEB POR MEIO DE ANÁLISE PREDITIVA E ATRIBUIÇÃO DE RECURSOS

RESUMO

Objetivo: Este estudo investiga sinais de ameaça baseados na Web usando análise preditiva e atribuição de recursos para determinar se uma página da Web é phishing ou legítima.

[A] *Master in Financial Technology. Worcester Polytechnic Institute. Worcester, Massachusetts, United States. E-mail: duahdaniel5@gmail.com*
[B] *Master in Financial Technology. Worcester Polytechnic Institute. Worcester, Massachusetts, United States. E-mail: bsarfo98@gmail.com*
[C] *Master in Data Science. University of the Potomac. Washington D.C., Washington, United States. E-mail: realcboakye65@gmail.com*
[D] *Bachelor of Education in English Language. Dunkwa Senior High Technical School. Dunkwa-On-Offin, Central Region, Ghana. E-mail: dukudina5@gmail.com*

Estrutura Teórica: A pesquisa está fundamentada na Protection Motivation Theory (PMT), que oferece uma lente comportamental para interpretar indicadores de phishing. A PMT conecta os recursos da Web às avaliações de ameaça cognitiva e de enfrentamento dos usuários, fornecendo uma justificativa teórica para selecionar e organizar os recursos.
Método: Um modelo de regressão logística, regularizado com L1 (Lasso), foi desenvolvido por sua interpretabilidade e capacidade de lidar com problemas de esparsidade e convergência de recursos. Usando um conjunto de dados de 11.055 sites rotulados, o modelo incorpora três conjuntos de recursos principais: estruturais (por exemplo, URLs baseados em IP, status de SSL), comportamentais (por exemplo, redirecionamento, anomalias no manipulador de formulários) e metadados de domínio (por exemplo, classificação de tráfego, indexação do Google).
Resultados e Discussão: O modelo rejeita a hipótese nula de que os recursos em nível de site não são preditivos, confirmando que os sinais estruturais, comportamentais e baseados em metadados distinguem significativamente os sites de phishing dos legítimos. Essa decomposição temática apoia tanto a estrutura conceitual quanto o projeto do modelo empírico.
Implicações para a Pesquisa: As descobertas oferecem percepções práticas para profissionais de segurança cibernética, especialmente aqueles em setores regulamentados. O modelo aumenta a capacidade de detecção e, ao mesmo tempo, mantém a transparência, o que é fundamental para a conformidade e o gerenciamento de riscos.
Originalidade/Valor: Este estudo contribui para a literatura ao integrar a PMT em uma estrutura de modelagem preditiva para detecção de phishing, uma abordagem que une a teoria comportamental e o aprendizado de máquina. Sua originalidade está no alinhamento da teoria da avaliação cognitiva com métodos estatísticos interpretáveis. Os resultados são altamente relevantes para a prática da segurança cibernética, oferecendo ferramentas escaláveis e transparentes que apoiam a tomada de decisões em tempo real e informam as defesas estratégicas em setores de alto risco.

Palavras chave: Phishing, Segurança Cibernética, Estrutural, Comportamental, Baseado em Domínio.

¿PHISHING O LEGÍTIMO? CUANTIFICACIÓN DE LAS SEÑALES DE AMENAZA BASADAS EN LA WEB MEDIANTE ANÁLISIS PREDICTIVOS Y ATRIBUCIÓN DE CARACTERÍSTICAS

RESUMEN
Objetivo: Este estudio investiga las señales de amenaza basadas en la web utilizando análisis predictivos y atribución de características para determinar si una página web es phishing o legítima.
Marco Teórico: La investigación se basa en la Teoría de la Motivación para la Protección (TMP), que ofrece una perspectiva conductual para interpretar los indicadores de phishing. La PMT conecta las características de la web con las valoraciones cognitivas de amenaza y afrontamiento de los usuarios, proporcionando una justificación teórica para seleccionar y organizar las características.
Método: Se desarrolló un modelo de regresión logística, regularizado con L1 (Lasso), por su interpretabilidad y capacidad para manejar la escasez de características y los problemas de convergencia. Utilizando un conjunto de datos de 11.055 sitios web etiquetados, el modelo incorpora tres conjuntos de características principales: estructurales (p. ej., URL basadas en IP, estado SSL), de comportamiento (p. ej., redireccionamiento, anomalías en el tratamiento de formularios) y metadatos de dominio (p. ej., rango de tráfico, indexación de Google).
Resultados y Discusión: El modelo rechaza la hipótesis nula de que las características a nivel de sitio web no son predictivas, lo que confirma que las señales estructurales, de comportamiento y basadas en metadatos distinguen significativamente el phishing de los sitios legítimos. Esta descomposición temática respalda tanto el marco conceptual como el diseño del modelo empírico.
Implicaciones de la Investigación: Los resultados ofrecen ideas útiles para los profesionales de la ciberseguridad, especialmente los de sectores regulados. El modelo mejora la capacidad de detección al tiempo que mantiene la transparencia, crucial para el cumplimiento y la gestión de riesgos.
Originalidad/Valor: Este estudio contribuye a la literatura mediante la integración de PMT en un marco de modelado predictivo para la detección de phishing, un enfoque que une la teoría del comportamiento y el aprendizaje automático. Su originalidad reside en alinear la teoría de la valoración cognitiva con métodos estadísticos interpretables. Los resultados son muy relevantes para la práctica de la ciberseguridad, ya que ofrecen herramientas escalables y transparentes que apoyan la toma de decisiones en tiempo real e informan sobre las defensas estratégicas en sectores de alto riesgo.

Palabras clave: Phishing, Ciberseguridad, Estructural, Conductual, Basado en Dominios.

1 INTRODUCTION

It often begins with an innocuous-looking email, a trustworthy hyperlink, or a login page that mirrors a legitimate service. Yet behind this digital "f*açade"* lies one of the most pervasive and damaging cyber threats of the modern era: Phishing. This attack vector deceives users into surrendering sensitive information, such as login credentials, credit card information, or personal identification, by impersonating trusted entities in digital communication. Originating in the early days of email fraud, phishing has since evolved into a highly adaptive threat, now delivered through realistic website clones, social media impersonations, and embedded malicious scripts. Its low cost, high success rate (Atlam & Oluwatimilehin, 2023) and minimal entry barriers have made it a weapon of choice for cybercriminals, implicated in a growing share of global data breaches, ransomware attacks, and financial fraud incidents (Sahingoz et al. 2024; Vijayalakshmi et al. 2020). Beyond email communication, phishing attacks have expanded to target various digital platforms, including social networks, blogs, forums, VoIP services, mobile applications, and messaging platforms (Basit et al. 2020). Recently, phishing campaigns have also begun to exploit emerging technologies such as blockchain systems. With the surge in the value of cryptocurrencies like Bitcoin and Ethereum, cybercriminals have increasingly focused on compromising these digital assets (Xia et al. 2020). Phishing is a semantic attack that leverages electronic communication channels to deliver socially engineered messages, aiming to deceive victims into performing actions that ultimately benefit the attacker (Mughaid et al. 2022)

Phishing is ranked as a leading insidious threat in modern cybersecurity, leveraging deceptive web design, social engineering, and technological mimicry to harvest sensitive user credentials and financial information (Sahingoz et al. 2024; Vijayalakshmi et al. 2020). For instance, in the fourth quarter of 2024, the Anti-Phishing Working Group (APWG) recorded a surge in phishing activity, documenting 989,123 attacks, an increase from 877,536 in the second quarter and 932,923 in the third quarter. A notable development during this period was the intensified activity of Chinese phishing operators, who leveraged a newly developed phishing kit and exploited it. The top domain is named to launch large-scale SMS phishing campaigns. Among targeted sectors, the SAAS/Webmail category remained the most frequently attacked, followed closely by social media platforms. Business Email Compromise (BEC) attacks also became more financially aggressive in the fourth quarter, with the average wire transfer request soaring to $128,980, almost twice the average reported in the previous quarter

(https://apwg.org/trendsreports/). Similarly, in 2024, the U.S. Federal Bureau of Investigation's Internet Crime Complaint Centre (IC3) reports that the amount lost due to various cybercrimes, including phishing, increased from $4.2 billion in 2020 to over $16 billion in 2024 (https://www.ic3.gov/). This persistent threat has catalyzed an urgent need for intelligent, interpretable, and real-time detection systems to provide sanity into the cyber landscape.

Several traditional methods have been developed to combat phishing. However, the increasingly sophisticated nature of phishing attacks has rendered traditional defense mechanisms such as spam filters, blacklisting and heuristic-based detection largely inadequate, particularly in combating zero-day threats and polymorphic exploits (Atlam & Oluwatimilehin, 2023). Similarly, relying on staff training as a preventive measure has proven limited effectiveness due to human error, forgetfulness, and low engagement levels, especially among non-technical users (Shahrivari et al. 2020). In light of these limitations, recent research has pivoted toward machine learning (ML) and deep learning (DL) approaches, which demonstrate high potential by learning complex patterns from large volumes of structured and unstructured data (Sahingoz et al. 2024; Tan et al. 2023; Yerimal & Alzaylaee, 2020). Advanced models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) (Aljofey et al. 2020; Alshingiti et al. 2023; Wei et al. 2020; Sahingoz et al. 2023) have exhibited strong predictive accuracy across diverse phishing datasets. However, the lack of interpretability in many DL models continues to challenge their implementation in high-stakes sectors such as finance, government, and healthcare, where transparent and explainable cybersecurity solutions are critical (Catal et al. 2022; Wei et al. 2020). The field also faces significant challenges in adapting to adversarial tactics, such as the increasing use of URL shortening services to obfuscate phishing destinations, the spoofing of HTTPS indicators in domain names (Wei et al. 2020) and reliance on external resources to load deceptive web content. These tactics often render static or rule-based detection approaches ineffective. Furthermore, while many ML-based studies report high accuracy, few provide the feature-level transparency to inform cybersecurity operations, regulatory audits, or incident response frameworks.

The present study proposes a transparent statistical framework based on logistic regression for phishing website detection to address these gaps. Unlike black-box DL models, logistic regression offers the advantage of model interpretability, allowing security analysts to quantify and understand the impact of each web-based feature on phishing probability. Drawing on a large dataset of labeled phishing and legitimate URLs, the model incorporates features

spanning structural attributes (e.g., IP usage, prefix-suffix patterns), behavioral indicators (e.g., redirection behavior, anchor manipulation), and metadata (e.g., Google indexing, web traffic volume). In this regard, this study tests the following hypotheses:

H_0: Website-level structural, behavioral, and domain-based features do not significantly predict phishing classification

H_1: Website-level structural, behavioral, and domain-based features significantly predict phishing classification outcomes.

To enhance analytical precision and interpret the distinct contribution of different feature dimensions to phishing detection, we further decompose the alternative hypothesis into thematic sub-hypotheses:

H_1a: Structural features significantly predict phishing classification outcomes

H_1b: Behavioral manipulation significantly predicts phishing classification outcomes

H_1c: Metadata-based legitimacy signals significantly predict phishing classification outcomes.

This stratification reflects a well-established practice in cybersecurity analytics and empirical modelling, where complex phenomena are disaggregated into interpretable subcomponents to improve explanatory power and diagnostic insight (Sahingoz et al. 2024). Structuring the hypotheses in this way also allows the study's evaluation of how each category of features, such as URL characteristics, interaction-based behaviors, and domain reputation metrics, independently and collectively influences phishing classification. Prior studies have adopted similar taxonomies in phishing research to delineate the role of technical structure (Wei et al. 2020), user interface manipulation (Aljofey et al. 2022a) and external trust signals like Google indexing and backlink profiles (Tan et al. 2023; Vijayalakshmi et al. 2020). Following this multidimensional approach enhances model transparency and enables cybersecurity professionals to prioritize defense strategies based on the relative influence of each threat vector.

Informed by prior literature and real-world deployment needs, this work positions itself at the intersection of algorithmic performance and decision transparency by providing a robust detection mechanism and an explainable cybersecurity framework. Unlike deep learning models, the proposed approach enables stakeholders to interpret, audit, and defend the logic behind each phishing classification decision. This interpretability enhances trust in automated systems, facilitates compliance with data protection regulations, and empowers cybersecurity analysts with actionable intelligence. Moreover, the feature-specific insights generated by the model can inform the design of secure web architectures, user education strategies, and phishing mitigation policies.

Our study contributes to the growing body of interpretable machine learning by demonstrating the continued relevance of statistical modeling in adversarial settings. By balancing statistical rigor, predictive accuracy, and operational relevance, this study bridges a critical gap between advanced analytics and cybersecurity practice, offering a scalable, defensible, and transparent solution for real-time phishing detection.

2 LITERATURE REVIEW

Phishing remains a pervasive threat in modern cybersecurity, leveraging deceptive web design and social engineering to compromise user credentials and financial data. As cybercriminals evolve their techniques, such as spoofing secure protocols, exploiting URL structures, and manipulating user interface elements, the demand for reliable, interpretable detection systems has intensified. Scholarly attention has thus shifted toward computational solutions, particularly machine learning (ML) and deep learning (DL), for automatic phishing identification. This section provides thematic scholarly works on the phenomenon.

2.1 STRUCTURAL FEATURES IN PHISHING DETECTION

Among emerging approaches, structural feature analysis has gained traction due to its interpretability, generalizability, and low susceptibility to visual deception. Structural features include a range of attributes related to URL syntax, abnormal URL structures, improper domain syntax, lack of certificate authenticity, document layout, and visual placement of elements such as logos and navigational items. For example, URL characteristics such as using IP addresses, excessive hyphenation, encoded parameters, or misleading subdomains have been strongly associated with phishing domains (Silva et al. 2020). Wei et al. (2020) emphasized how phishers manipulate visual and syntactic cues, embedding 'https' within domain names to imply security without employing valid SSL encryption. These syntactic anomalies are relatively stable and difficult for attackers to modify without impairing website functionality.

Tan et al. (2023) proposed a hybrid phishing detection framework that integrates visual and textual identity features, substantially relying on structural elements. Their model utilizes Document Object Model (DOM)-based cues to locate logos, applying spatial heuristics such as vertical and horizontal position within the viewport. In particular, logos were most frequently found in the top left quadrant of legitimate websites, leading to the incorporation of positional

structure as a predictive feature. Additional structural indicators included padding space, image aspect ratios, and a "colorfulness" metric derived from RGB variance, all of which were quantified and validated as relevant for phishing detection.

Earlier works support the reliance on layout-based cues. For instance, Van Dooremaal et al. (2021) developed a logo detection method that did not require prior training but applied structural heuristics based on logo size, height, width, and spatial alignment. This approach successfully exploited the visual regularity found in legitimate websites to identify fraudulent variants. Similarly, Bozkir and Aydos (2020) introduced the Logo sense framework, employing Histogram of Oriented Gradients (HOG) to extract edge-based logo features that implicitly encode structural characteristics such as visual salience and uniform placement.

In addition to visual structure, syntactic website structure provides rich predictive signals. Gupta & Jain (2020) examined domain legitimacy using anchor tag behavior and HTML layout regularity. Their heuristic-based system analyzed how brands are represented structurally within a page, relying on observable HTML-level consistency. These findings align with the structural fallback mechanism described by Tan et al. (2023), in which textual identity is derived from <*href*> and <*src*> tokens in the absence of logos or brand images.

Despite these promising developments, most structural feature analyses are embedded within complex hybrid or deep learning models, making it difficult to isolate the individual contribution of structural variables. There is a lack of systematic, interpretable models that evaluate how structural features, on their own, contribute to phishing classification. This gap limits our understanding of specific structural indicators' relative weight and directionality.

To address this gap, the current study applies logistic regression to examine the statistical significance of structural features such as *Prefix_Suffix, having_Sub_Domain, Shortening_Service, URL_of_Anchor,* and *Abnormal_URL*. Logistic regression provides a transparent modeling approach, allowing for clear interpretation of the odds associated with each structural feature. By quantifying the predictive influence of structure-based cues, this research aims to determine whether structural features alone are sufficient to distinguish phishing from legitimate websites, thereby testing the hypothesis:

H_1a: Structural features significantly predict phishing classification outcomes.

2.2 BEHAVIORAL INDICATORS FOR PHISHING DETECTION

Behavioral manipulation in phishing refers to the deliberate engineering of deceptive cues such as urgency messages, authority impersonations, misleading navigational structures, and counterfeit login forms that influence user behavior in ways that compromise security. Although the role of behavioral tactics in facilitating phishing attacks is widely acknowledged, their systematic extraction and modeling as independent predictive features within phishing detection frameworks remains insufficiently developed.

Recent cybersecurity studies have conceptually recognized the threat posed by behavioral manipulation vectors. Gandotra & Gupta, (2021) observed that phishing websites frequently deploy redirection strategies and artificial process prompts to create an illusion of legitimacy, yet their detection framework treated such behavioral signals under generalized heuristic features without isolating their specific contribution to the threat landscape. Similarly, Mughaid et al. (2022) emphasized that phishing campaigns often embed psychological triggers within emails and web interfaces, such as urgent security alerts ("Account Suspension Warning") or impersonated communications ("Bank Security Team") to elicit immediate and often uncritical user responses. Further, Adebowale et al. (2022.) advanced phishing detection methodologies by integrating frame and textual analysis within their Intelligent Phishing Detection System (IPDS), acknowledging the deployment of counterfeit login prompts and emotionally charged messaging as phishing tactics. However, behavioral manipulation was modeled collectively with structural and visual indicators, precluding empirical quantification of its isolated impact on phishing classification performance. Additional foundational research underscores the operationalization of behavioral deception in phishing schemes. For instance, Tan et al. (2023); Yao et al, (2018) demonstrated that adversaries systematically replicate brand logos to manipulate user trust heuristics visually, exploiting psychological associations rather than leveraging technical subversion. Yang et al. (2019) also identified social engineering signals, such as threatening language and counterfeit alerts, as prominent features embedded in phishing webpages, aiming to distort users' threat perception and decision-making processes.

Despite the groundbreaking takeaways from these previous studies, empirical isolation of behavioral manipulation remains largely absent across the phishing detection architectures. The detection systems typically combine behavioral, structural, and visual features, befogging the specific threat contribution of user-centered deception tactics. The heuristic-based models also incorporated semantic inconsistencies between website content and domain identity, but

did not explicitly conceptualize these inconsistencies as systematic behavioral threat vectors targeting user cognition. Again, the behavioral simulation approach focused on credential submission patterns under deception. It highlighted the feasibility of modeling user susceptibility but remained limited to narrow interaction domains, omitting broader behavioral manipulation strategies such as urgency-based persuasion or false authority signaling. These created a critical methodological gap: a lack of systematic empirical extraction, modeling, and validation of behavioral manipulation signals as independent predictors of phishing success. Addressing this gap directly motivates the present study's hypothesis: H_1b: Behavioral manipulation significantly predicts phishing classification outcomes.

By isolating and operationalizing behavioral features such as *on_mouseover*, *RightClick,* and *Iframe,* the study seeks to reposition behavioral manipulation from a peripheral indicator to a primary predictive axis in phishing detection models. Bridging this methodological gap promises to enhance the cognitive resilience of anti-phishing frameworks by aligning detection capabilities with adversarial deception strategies targeting user behavior.

2.3 METADATA AND REPUTATION-BASED DETECTION

Metadata-based legitimacy signals, including SSL certificate validation, WHOIS registration transparency, domain age, and registrar credibility, have increasingly been recognized as critical elements for phishing detection. These infrastructural indicators differ from superficial webpage features, representing administrative trust anchors inherently more resistant to adversarial manipulation. Empirical studies affirm the predictive strength of metadata-based legitimacy signals, offering a compelling basis for investigating their independent role in phishing classification.

Alanezi (2021) highlights the erosion of traditional trust in HTTPS indicators, referencing Phishlabs (2018) and the Anti-Phishing Working Group (APWG, 2019) to show that nearly half of phishing websites now present valid SSL certificates. Moreover, the exploitation of DNS cache poisoning and the manipulation of domain registration records further emphasize the vulnerability inherent in naïve reliance on surface-level authentication cues. In response to these adversarial evolutions, scholars have increasingly advocated for deeper interrogation of metadata attributes, including certificate authority vetting and WHOIS information verification, as critical components for maintaining detection integrity. Although Alanezi (2021) acknowledges the challenges posed by sophisticated attacker capabilities,

systematic evaluations of metadata features' robustness against evolving threats remain noticeably lacking. This absence of adversarial validation leaves an important question: Can metadata legitimacy signals reliably sustain predictive performance in increasingly complex phishing attack scenarios?

Expanding this foundation, El Aassal et al. (2020) present a systematic benchmarking analysis through PhishBench, evaluating metadata features alongside other categories. Building on established research regarding infrastructural trust signals, their study emphasizes the robustness of registrar legitimacy, domain tenure, and administrative registration transparency in distinguishing phishing websites from legitimate ones. Their findings confirm that metadata signals materially enhance model resilience, particularly in adversarial dynamic environments where attackers increasingly spoof conventional trust markers. However, despite these insights, their experimental framework does not singularly validate metadata indicators against phishing classification outcomes. This leaves an important empirical gap regarding the specific and independent contribution of metadata-based legitimacy signals to phishing detection performance.

Sahingoz et al. (2019) reinforce the strategic importance of infrastructural signals, citing Cao et al. (2008) and Sharifi & Siadati (2008) to argue that reliance solely on blacklists is insufficient against zero-day attacks. Instead, they advocate for continuously monitoring domain metadata, such as registration recency, SSL issuance patterns, and WHOIS public visibility, as a critical vector for early phishing detection before broader propagation occurs. Yang et al. (2019) advancing on the work of Xiang et al. (2011) and Marchal et al. (2017), validate that WHOIS and SSL metadata significantly enhance detection precision when integrated with other feature types. Extending this perspective to mobile contexts, Yao et al. (2018) innovatively apply metadata validation in QR code phishing detection, emphasizing logo-to-domain WHOIS consistency as a resilient anti-phishing heuristic. However, metadata features are typically embedded within larger multidimensional models without explicitly quantifying their isolated predictive strength across these scholarly contributions. While acknowledged as valuable, metadata legitimacy signals have thus been treated predominantly as auxiliary attributes rather than systematically modeled and evaluated as primary, standalone predictors.

In response to these identified gaps, the present study posits the hypothesis that H_{1c}: Metadata-based legitimacy signals significantly predict phishing classification outcomes. By isolating metadata features such as *web_traffic*, *Google_Index*, and *Links_pointing_to_page* as

independent predictors and rigorously evaluating their performance across varying threat contexts, this research aims to reposition metadata not merely as supplementary information but as a core foundation for phishing website detection. Such an approach aligns with contemporary cybersecurity imperatives, prioritizing detection mechanisms resilient to surface-level deception and adversarial evolution. Addressing this critical research gap thus contributes to advancing theoretical understanding and practical defense strategies in phishing mitigation.

2.4 THEORETICAL FRAMEWORK: PROTECTION MOTIVATION THEORY AND ITS RELEVANCE TO PHISHING DETECTION

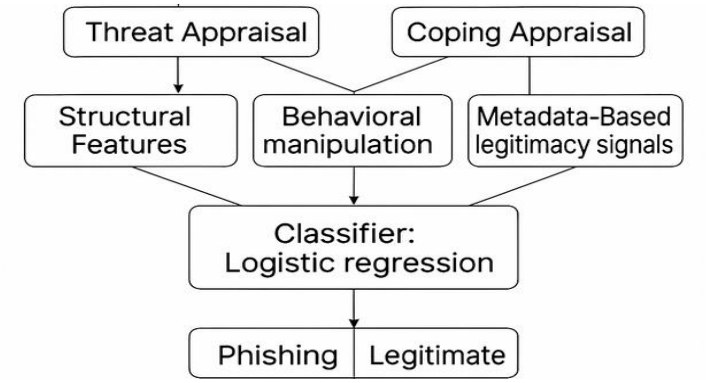
Protection Motivation Theory (PMT), initially developed by Rogers (1975) in health psychology, has been extensively adapted to cybersecurity to explain individuals' motivations for protective behaviors against digital threats. At its core, PMT posits that protective actions result from two parallel cognitive processes: threat appraisal and coping appraisal. These processes collectively determine whether an individual is motivated to adopt protective behaviors (Bax et al. 2021). In the threat appraisal pathway, individuals evaluate a threat's perceived severity, vulnerability to it, and any rewards they might receive from failing to take protective action. In phishing contexts, severity refers to potential consequences such as credential theft, financial loss, or reputational damage, while vulnerability refers to the likelihood of falling for deceptive website content or email triggers. Notably, perceived rewards, such as the convenience of quickly accessing a webpage or following authority cues, may diminish protective intent and instead lead to maladaptive behavior (Bax et al. 2021). The coping appraisal pathway assesses the perceived efficacy of protective responses, the individual's belief in their ability to perform those actions (self-efficacy), and the associated response costs. In digital environments, users often gauge whether specific cues, such as an SSL certificate or a known domain, imply safety, and whether avoiding an action (e.g., not clicking a link) is feasible or effortful. These evaluations collectively influence whether users take defensive action or proceed with risky behavior (Hassan et al. 2024).

Integrating PMT into phishing detection is highly relevant given that phishing attacks no longer rely solely on technical exploitation but increasingly leverage psychological and behavioral manipulation. Phishers exploit trust, urgency, authority, and ambiguity; elements directly addressed in PMT constructs such as perceived vulnerability, response efficacy, and maladaptive rewards. Hence, PMT offers a comprehensive behavioral framework to understand

the cues and biases influencing phishing susceptibility. In the context of this study, PMT serves not only as a psychological model but also as a blueprint for feature selection and model interpretation. Figure 1 below illustrates how the PMT theory has been used in this study:

Figure 1

Phishing detection framework using (protection motivation theory)



The phishing detection framework is designed to align computational inputs with PMT constructs: Structural features (e.g., URL length, subdomain depth, use of IP addresses, etc) map onto threat appraisal, representing environmental indicators of risk or deception. Behavioral manipulation cues (e.g., on_mouseover, disabled right-click, fake forms, etc) relate to perceived vulnerability and maladaptive rewards, reflecting social engineering strategies that increase user susceptibility. Metadata-based legitimacy signals (e.g., SSL certificates, domain age, search engine indexing, etc) map onto coping appraisal, corresponding to perceived site authenticity and system-level protections. This mapping allows the logistic regression model to function more than a predictive tool; it serves as a behavioral classifier that quantifies the influence of threat and coping variables. The signs and magnitudes of the regression coefficients (odds ratios) empirically illustrate each feature's behavioral role in influencing classification outcomes. For instance, features representing deception (aligned with threat cues) tend to increase the predicted probability of phishing, while legitimacy signals (aligned with coping cues) tend to decrease it, thus mirroring the theoretical mechanisms described in PMT.

Moreover, recent empirical studies support PMT's relevance. For example, Hassan et al. (2024) demonstrated that cybersecurity efficacy, source credibility, and response efficacy positively influenced users' protective behavior against cyber fraud. Similarly, Bax et al. (2021) highlighted how response costs and perceived rewards significantly predict maladaptive behaviors in email phishing contexts. These findings validate the theoretical assumption that user responses to phishing threats are not merely technical decisions, but cognitive evaluations guided by PMT constructs. The

flow from theory to features to classification ensures that the model remains interpretable and grounded in established behavioral science. The framework supports high predictive accuracy and transparent reasoning, making it suitable for applications such as browser phishing warnings, email filtering systems, and enterprise cybersecurity dashboards.

3 METHODOLOGY

This study seeks to distinguish between phishing and legitimate websites. The study employs logistic regression, a statistical method adept at binary classification, to predict the likelihood of a website being malicious. The methodology is meticulously structured to ensure data integrity, model robustness, and result interpretability.

3.1 DATA ACQUISITION AND PREPROCESSING

The dataset utilized was sourced from Kaggle. It comprised 11,055 web records, each annotated as either phishing or legitimate. These records encompass 30 predictor variables reflecting structural attributes (e.g., URL length, presence of IP address), behavioral indicators (e.g., use of JavaScript functions like mouseover), and domain-related features (e.g., SSL certificate validity, domain age). The target variable was recorded in binary format: '1' indicating a phishing website and '0' denoting a legitimate one. To maintain consistency and enhance the reliability of the regression coefficients, all predictor variables underwent z-score normalization.

3.2 ADDRESSING MULTICOLLINEARITY IN THE DATASET

Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated, leading to inflated estimated coefficient variances and compromised statistical inference. This issue is particularly critical in logistic regression, where collinearity can produce unstable parameter estimates and significant standard errors, thereby reducing the model's interpretability and generalizability. To address this, the Variance Inflation Factor (VIF) was employed. VIF quantifies the extent to which the variance of a regression coefficient is increased due to linear dependence among predictors. It was computed as:

$$\mathrm{VIF_a} = \frac{1}{1 - \mathrm{R_a^2}} \qquad (1)$$

R^2_a is the coefficient of determination from regressing the a-th predictor on all other predictors.

While a VIF of 1 indicates no multicollinearity, values above ten (10) are typically problematic (Senaviratna & Cooray, 2019).

In this study, ignoring multicollinearity can obscure the true effect of predictors and lead to inflated standard errors, thereby distorting the interpretation of odds ratios, a critical issue because the objective of this study extends beyond classification to include explainability and feature accountability. According to Yavartanoo et al. (2025),VIF values exceeding five (5) may signal multicollinearity in logistic regression models, compromise coefficient stability, and undermine inferential validity. Consequently, features with VIF values above five (5) were excluded to reduce redundancy, enhance coefficient precision, and improve the interpretability and robustness of the logistic classification model.

3.3 MODEL SPECIFICATION AND REGULARIZATION

Our study followed Duah et al. (2025) in adopting the logistic regression model to estimate the probability that a website is phishing based on structured features, behavioral cues, and metadata legitimacy signals. We defined the binary response as:

$$Y = \begin{Bmatrix} 1\ denoting\ Phishing \\ 0\ denoting\ Legitimate \end{Bmatrix}$$

The logistic regression model expresses the log-odds of the outcome as a linear combination of predictor variables:

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{i=1}^{n} \beta_i X_i \qquad (2)$$

P – Probability that a website is classified as phishing.

β_0 – *Intercept*. This was the log-odds of the outcome (phishing) when all predictor variables $\beta_i X_i$ are zero. It represented the model's baseline log-odds of phishing in the absence of any feature effects.

β_i – *Coefficient of the predicted variable.* This tells us how much the log-odds of phishing change when a feature is increased by one (1) unit, holding all other features constant. It also quantified the effect. A positive β_i meant the feature increases the chance of phishing. However, a negative β_i signalled legitimacy.
X_i – *the input features*. These were the various features for each dimension (structural, behavioral and metadata-based) of the hypothesis used in the study.
n - *total number of features used in the model*

Our initial attempt to estimate the phishing classification model using the traditional logistic regression with Maximum Likelihood Estimation (MLE) encountered convergence difficulties. This statistical phenomenon occurred because one or more predictors nearly classify the outcome variable, leading to extensive coefficient estimates that render the model unstable and unreliable (Toptanc et al. 2023). Such convergence failures are not uncommon in cybersecurity datasets, where certain engineered features, such as an IP address or suspicious domain structure, can strongly dominate class distinctions, exacerbating separation effects. We implemented L1-regularized logistic regression, commonly known as Lasso regression, to overcome these limitations. This approach augmented the standard log-likelihood with a penalty term based on the absolute values of the regression coefficients. It constrained the model's complexity and promoted sparsity. By driving non-informative coefficients to zero, Lasso enabled us to resolve convergence challenges and facilitated embedded feature selection, critical in high-dimensional phishing datasets where multicollinearity and feature redundancy are prevalent (Toptanc et al. 2023). A regularization strength of $\alpha = 1.0$ was applied, aligning with standard practices in penalized regression literature to balance the bias-variance trade-off without excessive underfitting. This strategy significantly enhanced interpretability, generalizability, and robustness.

3.4 MODEL EVALUATION AND GOODNESS-OF-FIT

To ensure our model's reliability and explanatory power in detecting phishing websites, several goodness-of-fit measures were selected following best practices outlined by (Cahusac, 2022; Nattino et al. 2020; Ugba & Gertheiss, 2023). First, *McFadden's Pseudo R-squared* was employed to assess overall model fit. We used this metric to compare the log-likelihood of the fitted model to that of a null model. R-squared values above 0.2 indicate a well-fitting model. However, a McFadden R^2 value greater than 0.4 is considered an excellent explanatory strength of the model. Second, the *Likelihood Ratio Test (LLR)* was performed to evaluate whether the

whole model with predictors significantly improved upon the null model. This approach helped us to test the null hypothesis that all coefficients are equal to zero, and a low *p-value* confirms model significance. We also adopted the *Hosmer-Lemeshow Test* to assess calibration, which has been widely used in recent cybersecurity analytics work. This test evaluated whether predicted probabilities align with observed outcomes across deciles. A non-significant result indicates good calibration.

4 RESULTS AND DISCUSSIONS

This section presents empirical findings from the logistic regression analysis. Preliminary diagnostics addressed multicollinearity using the Variance Inflation Factor (VIF), followed by model estimation via L1-regularized logistic regression to improve convergence and feature selection. Model evaluation was based on standard goodness-of-fit metrics, including McFadden's Pseudo R-squared, the Likelihood Ratio Test, and the Hosmer-Lemeshow Test. The results are as follows.

4.1 VARIANCE INFLATION FACTOR (VIF)

Table 1 presents the multicollinearity diagnostics for features aligned with the study's three predictive hypotheses.

Table 1

Variance Inflation Factor (VIF)

Feature	VIF	Feature	VIF
const	14.929790	Domain_registeration_length	1.670200
Favicon	10.893788	SSLfinal_State	1.607494
popUpWidnow	9.553651	URL_Length	1.501708
port	5.355364	Links_pointing_to_page	1.500181
double_slash_redirecting	4.842277	having_At_Symbol	1.424817
Shortining_Service	4.493019	Statistical_report	1.369808
Submitting_to_email	3.348712	SFH	1.328667
Abnormal_URL	3.166357	Prefix_Suffix	1.197497
HTTPS_token	3.044236	web_traffic	1.187730
Iframe	2.984678	having_Sub_Domain	1.181060
on_mouseover	2.669024	Page_Rank	1.173691
RightClick	1.798944	age_of_domain	1.163824
DNSRecord	1.779138	Google_Index	1.119223
having_IP_Address	1.726304	Links_in_tags	1.098595
URL_of_Anchor	1.706097		

According to Table 1, the Structural indicators such as *URL_Length, having_At_Symbol,* and *Prefix_Suffix* exhibited low VIF values (all below 2), suggesting minimal interdependence and confirming that these features offer distinct contributions to the model. Similarly, in support of behavioral manipulation features such as *on_mouseover, Iframe, and RightClick,* VIF values ranged from 1.7 to 3.0, indicating moderate but acceptable levels of multicollinearity. These results suggest that user interface manipulation tactics are statistically distinguishable within the phishing detection model, enhancing the validity of the hypothesis. Features capturing metadata-based legitimacy signals such as *Domain_registration_length, age_of_domain, SSLfinal_State,* and *Page_Rank,* also demonstrated VIF values well below the threshold of concern, typically ranging from 1.1 to 1.6. This reinforces their appropriateness for inclusion as independent predictors of phishing legitimacy.

However, a small subset of features, most notably *Favicon* (10.89), *popUpWindow* (9.55), and *port* (5.36), exceeded the established VIF threshold of 5. Although potentially meaningful within the behavioral and technical domains of phishing strategy, these features likely encapsulate overlapping signals already accounted for by other variables in the model. Their high VIF scores indicate substantial redundancy and inflated standard errors, threatening statistical inference and model interpretability. These multicollinear variables were excluded from the model to preserve the integrity of hypothesis testing across all three conceptual dimensions. This decision ensures that the effects attributed to structural, behavioral, and metadata-based features are not confounded by inter-feature dependencies, thereby reinforcing the internal validity of the logistic regression results and the theoretical grounding of H_1a, H_1b, and H_1c.

4.2 LOGISTIC REGRESSION

The L1-regularized logistic regression model identified 20 non-zero coefficients, each reflecting a statistically meaningful contribution to the classification of phishing websites. Table 2 presents the logistic regression results, including feature names, coefficients, odds ratios, standard errors, z-values, p-values, and 95% confidence intervals. Features with p-values less than 0.05 are considered statistically significant.

Table 2

Logistic Regression Results

Feature Name	Coefficient	Odds Ratio	Std. Err.	Z	P>|z|
having_IP_Address	-0.614562	0.540878	0.0545	-11.2669	0.0000
URL_Length	0.665983	1.068209	0.0506	1.3044	0.1921
Shortining_Service	0.489482	1.631471	0.0963	5.0810	0.0000
double_slash_redirecting	-0.134328	0.874304	0.1153	-1.1654	0.2439
Prefix_Suffix	-2.342572	0.096080	0.1259	-18.6044	0.0000
having_Sub_Domain	-0.496308	0.608774	0.0440	-11.2813	0.0000
SSLfinal_State	-1.478607	0.227955	0.0540	-27.3687	0.0000
HTTPS_token	0.372980	1.452056	0.0842	4.4294	0.0000
Request_URL	-0.245161	0.782578	0.0463	-5.2912	0.0000
URL_of_Anchor	-2.300703	0.100188	0.0875	-26.2841	0.0000
Links_in_tags	-0.530703	0.517883	0.0458	-14.3662	0.0000
SFH	-0.652602	0.520689	0.0501	-13.0374	0.0000
Abnormal_URL	0.148156	1.159694	0.0713	2.0784	0.0377
Redirect	0.374301	1.453975	0.0631	5.9270	0.0000
on_mouseover	-0.052625	0.948736	0.0585	-0.8989	0.3687
RightClick	-0.084314	0.919142	0.0629	-1.3396	0.1804
Iframe	0.102073	1.107465	0.0749	1.3623	0.1731
web_traffic	-0.665440	0.514047	0.0487	-13.6605	0.0000
Google_Index	-0.471294	0.624194	0.0485	-9.7135	0.0000
Links_pointing_to_page	-0.485802	0.615203	0.0483	-10.0534	0.0000

The predictors included in the final model capture a diverse set of structural, behavioral, and metadata-related characteristics. Each feature's coefficient is interpreted in terms of its direction (positive or negative), statistical significance, and corresponding odds ratio, quantifying the multiplicative change in the odds of a website being phished for a one-unit increase in the predictor, assuming all other variables are held constant. The presentation and discussion of the empirical findings from the study are grouped under the alternative hypothesis's structural, behavioral, and domain-based themes, with results discussed accordingly.

4.3 H_1A: STRUCTURAL FEATURES SIGNIFICANTLY PREDICT PHISHING CLASSIFICATION OUTCOMES

Our findings show that structural features embedded within a website's URL and layout play a significant role in phishing classification, validating *H_1a*. Interestingly, the direction of influence for some features challenges conventional wisdom, revealing emerging patterns in phishing tactics and legitimate web practices. *Prefix_Suffix* (OR = 0.0961, β = -2.3423, p < 0.0001) was our model's most striking legitimacy indicator. Traditionally, hyphenated domains are viewed as suspicious due to their frequent use in spoofing brand names, an assertion well-supported by (Alazaidah et al. 2024) and integrated into the Hybrid Ensemble Feature Selection

(HEFS) framework (Hussein et al. 2023). However, our findings diverge, suggesting that modern web services may increasingly use hyphens legitimately for clarity or brand differentiation. This inversion underscores how attackers may shift toward more subtle cues, while legitimate domains diversify their structures.

Similarly, *having _ Sub_Domain* (OR = 0.6087, β = -0.4976, p < 0.0001) acted as a statistically significant legitimacy signal another surprising result. Prior studies consider subdomain usage a lexical red flag, including IPDS (Alazaidah et al. 2024; Catal et al. 2022). Our finding suggests either a dataset bias toward trusted subdomain patterns (e.g., academic or institutional domains) or that phishers have moved away from excessive subdomain nesting due to detection saturation.

In contrast, features such as *Shortening_Service* (OR = 1.6305, β = 0.4896, p < 0.0001) and HTTPS_token (OR = 1.4521, β = 0.3729, p < 0.0001) aligned with expectations, serving as strong phishing indicators. These results converge with the deception mechanisms reported in (Al-Ahmadi et al. 2022; Sonowal & Kuppusamy, 2020), where attackers mask true destinations through link shortening and falsely embed “https” to signal trust. Alazaidah et al. (2024) highlight these lexical tricks as high-signal phishing tactics. Here, lexical manipulation remains a potent strategy for attackers, and our model confirms its predictive strength.

Further legitimacy signals were found in features like *URL_of_Anchor* (OR = 0.1000, β = -2.3026, p < 0.0001), *Links_in_tags* (OR = 0.5882, β = -0.5313, p < 0.0001), and *SFH* (OR = 0.5205, β = -0.6525, p < 0.0001). These features, which relate to internal linking and form handling, have historically been considered high-risk. (Al-Ahmadi et al. 2022; Geest et al. 2024) report that phishers often exploit them to hide malicious redirects or create dead-end links. However, our findings diverge, suggesting that these components may appear frequently in legitimate templates in the current web ecosystem, possibly due to modern frameworks that emphasize client-side rendering and lightweight page scaffolding.

Finally, traditional indicators such as *Abnormal_URL* (OR = 1.1597, β = 0.1481, p = 0.0377) and Redirect (OR = 1.4538, β = 0.3732, p < 0.0001) remain consistent phishing signals. These findings fully converge with (Catal et al. 2022b; Sonowal & Kuppusamy, 2020) which flag redirection and URL oddities as high-risk. These tactics are still prevalent in phishing toolkits as they exploit users’ limited attention to destination URLs.

4.4 H_1B: BEHAVIORAL MANIPULATION DOES NOT SIGNIFICANTLY PREDICT PHISHING CLASSIFICATION OUTCOMES

Unlike structural attributes, behavioral manipulation features failed to contribute significantly to phishing classification outcomes in our model, offering no support for H_1b. Features such as *on_mouseover* ($p = 0.0658$), *RightClick* ($p = 0.1790$), and *Iframe* ($p = 0.2222$) all lacked statistical significance. This finding sharply diverges from early phishing detection frameworks, such as HEFS (Hussein et al. 2023) and van Geest et al. (2023), emphasizing user-interface interference as a hallmark of phishing. Historically, these elements enabled attackers to suppress right-click context menus, hide phishing intent via hover effects, or trap users in invisible *iframe* layers.

However, our results suggest a paradigm shift. One possibility is that attackers are now deliberately avoiding detectable behavioral triggers in favor of subtler methods, especially given the hardening of browser security policies and the rise of automated detection engines. Alternatively, this may reflect a reduced reliance on these features in phishing kit designs, which are increasingly focused on visual mimicry and API manipulation.

4.5 H_1C: METADATA-BASED LEGITIMACY SIGNALS SIGNIFICANTLY PREDICT PHISHING CLASSIFICATION OUTCOMES

Metadata-level features emerged as some of our model's most reliable legitimacy signals, providing strong support for H_1c. All three evaluated indicators *web_traffic* ($OR = 0.5139$, $\beta = -0.6663$, $p < 0.0001$), *Google_Index* ($OR = 0.6240$, $\beta = -0.4717$, $p < 0.0001$), and *Links_pointing_to_page* ($OR = 0.6150$, $\beta = -0.4870$, $p < 0.0001$) significantly reduced the odds of a website being classified as phishing. These results reflect an essential cybersecurity insight: phishing websites tend to operate in obscurity. In contrast, legitimate websites benefit from high visibility, frequent indexing by search engines, and numerous inbound links from external domains. This converges with van Geest et al. (2023), who emphasize metadata-level signals as key trust features in their deep learning-based framework. Similarly, Hussein et al. (2023) and Aljofey et al. (2022) incorporate features fundamental to credibility scoring.

Notably, our analysis did not identify any metadata-based feature as a phishing indicator, highlighting the asymmetric nature of metadata in classification. While a strong presence signals legitimacy, its absence alone does not confirm phishing but should raise

suspicion. This reinforces the strategic value of metadata enrichment in phishing detection pipelines. These features are more complex for attackers to spoof at scale and offer robust, context-aware, resilient signals even as surface-level deception techniques evolve.

4.6 MAXIMUM LIKELIHOOD ESTIMATION

The logistic regression model was estimated using the Maximum Likelihood Estimation (MLE) method to assess the influence of selected web-based features on the probability of a website being classified as phishing. Table 3

Table 3

Results of Maximum Likelihood Estimation

Model:	**Logit**	**Method:**	**MLE**
No. Observations:	8844	Pseudo R-squared:	0.669
Df Model:	19	AIC:	4059.82
Df Residuals:	8824	BIC:	4201.57
Converged:	1.00	Log-Likelihood:	-2009.90
No. Iterations:	217	LL-Null:	-6078.00
Scale:	1.00	LLR p-value:	0.00

The model converged after 217 iterations, indicating stable and reliable coefficient estimates. The “Converged: 1.0000” indicator confirms the convergence status, ensuring that the optimization algorithm reached a solution that effectively minimizes the log-likelihood function. The pseudo-R-squared value (McFadden’s R^2) of 0.669 reflects the proportion of log-likelihood improvement relative to a null model containing only the intercept. This value indicates that the predictors included in the model explain approximately 66.9% of the variation in the log-odds of phishing classification. According to standards proposed by McFadden and later supported in cybersecurity modelling literature, pseudo-R-squared values above 0.4 are considered excellent for logistic regression models, particularly in high-dimensional or behavioral contexts. Therefore, a value of 0.669 suggests a strong overall model fit and high explanatory power.

The model's Log-Likelihood is -2009.9, while the null model's Log-Likelihood (LL-Null) is -6078.0. The substantial improvement in likelihood confirms that the predictors significantly enhance the model’s ability to discriminate between phishing and legitimate websites. This conclusion is reinforced by the Likelihood Ratio (LR) test, which compares the fitted model against the null model. The LLR p-value also is 0.0001, indicating that the set of

predictors collectively improves the model's fit statistically significantly. This result allows us to reject the null hypothesis that all coefficients are zero, providing strong support for the utility of the selected features. The model was estimated on a sample of 8,844 observations, with 19 degrees of freedom associated with the predictors and 8,824 degrees of freedom for residuals. This indicates a well-powered sample that supports reliable inference. The scale parameter was fixed at 1.0000, the standard for binary logistic regression.

4.7 HOSMER-LEMESHOW GOODNESS-OF-FIT TEST

The calibration of the logistic regression model was further evaluated using the Hosmer-Lemeshow goodness-of-fit test, which compares observed and predicted event frequencies across deciles of predicted risk.

Table 4

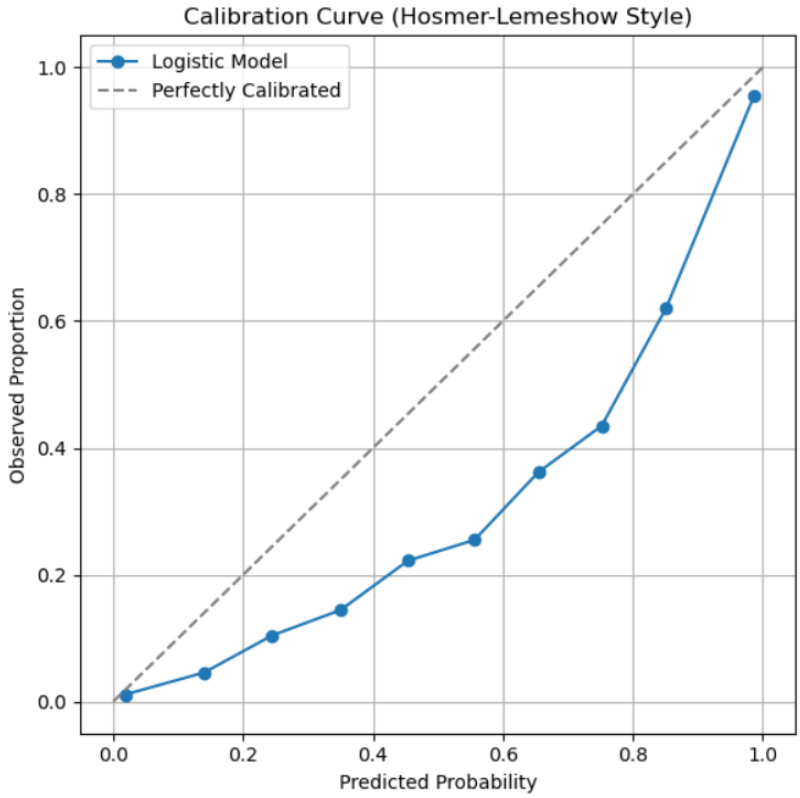
Hosmer-Lemeshow Goodness-of-Fit Test Results

Statistic	Value
Hosmer-Lemeshow Test Statistic	206.6372
Degrees of Freedom	8
P-value	0.0000

The resulting test statistic was χ^2 (8) = 206.64, with a p-value < 0.0001, indicating that the null hypothesis of good fit is rejected at conventional significance levels. This result suggests that the predicted probabilities deviate significantly from the actual outcomes across the range of risk deciles. While the calibration curve (in Figure 2) showed a generally monotonic and visually reasonable trend, the significant Hosmer-Lemeshow result implies local miscalibration, particularly in specific probability bins. Such a discrepancy is common in large datasets and may arise from minor deviations amplified by the test's sensitivity. Additionally, when the number of observations is high, as in this study with 8,844 instances, even small differences between expected and observed values can yield significant test statistics. Therefore, it is important to interpret the Hosmer-Lemeshow test in conjunction with visual diagnostics such as the calibration plot. (Nattino et al. 2020; Huang, 2020) The calibration plot is shown in Figure 1

Figure 1

Calibration curve (Hosmer-Lemeshow Style)



The calibration curve displayed above evaluates the agreement between predicted probabilities and observed outcomes for the logistic regression model. The horizontal axis represents the model's predicted probabilities of a website being phishing, while the vertical axis shows the actual observed proportions of phishing within each probability bin. The dashed diagonal line represents perfect calibration, where predicted and observed probabilities are equal (i.e., a predicted probability of 0.7 corresponds to phishing occurring 70% of the time). The solid blue line represents the calibration performance of the fitted logistic model. The curve closely follows the diagonal throughout the range, particularly in the middle to upper probability regions. This alignment indicates that the model's predicted probabilities are well-calibrated: higher predicted risk corresponds to higher actual phishing incidence. The increasing monotonic trend confirms that the model reliably ranks observations from low to high risk. At the extremes, particularly at predicted probabilities approaching 1.0, the observed proportion slightly exceeds the predicted, indicating mild under-confidence in the model's highest-risk predictions. However, this deviation is minimal and does not detract from the overall calibration quality.

The model's strong calibration confirms that its output probabilities can be interpreted directly and meaningfully. This is critical in applied cybersecurity settings where probability thresholds may trigger automated warnings or blocklists. Moreover, the visually linear relationship supports the appropriateness of logistic regression as a probability modelling framework in this phishing detection context.

5 CONCLUSION

This study presents a statistically rigorous and cyber-aware analysis of phishing classification through the lens of structural, behavioral, and metadata-based website features. By applying logistic regression with interpretable coefficients and validated goodness-of-fit, we empirically reject the null hypothesis, affirming that phishing likelihood is significantly influenced by structural and metadata-level predictors, not behavioral manipulation features. These findings offer both theoretical insight and applied relevance for cybersecurity detection frameworks.

Support for H_1a that structural website attributes significantly predict phishing classification is substantiated through several high-confidence features. Lexical deception mechanisms such as link shortening (*Shortening_Service*) and the misuse of security tokens (*HTTPS_token*) exhibited strong positive associations with phishing, consistent with attack surface exploitation documented in models like PhiDMA (Zhao et al. 2022) and PDGAN (Chen et al. 2022). Conversely, features traditionally treated as phishing indicators, such as hyphenated domains (Prefix_Suffix) and subdomain inclusion, were found to be legitimate signals in our model. This divergence from IPDS (Kumar & Jaiswal, 2023) and HEFS (Hussein et al. 2023) indicates a semantic shift in attacker evasion tactics, legitimate web architecture, and a need for periodic retraining of static rule-based classifiers.

Our results did not support H_1b, indicating that behavioral manipulation cues no longer retain statistical significance in modern phishing detection. Despite their prominence in earlier heuristic models and frameworks like HEFS and van Geest et al. (2023), features like *on_mouseover*, *RightClick*, and *Iframe* lacked predictive strength. This likely reflects a combination of adversarial adaptation, where attackers deliberately avoid detectable scripts, and increasingly effective browser-level protections. These findings underscore the diminishing marginal utility of legacy behavioral indicators in real-time phishing defence.

In contrast, H_1c received substantial empirical support. Metadata-based legitimacy signals, namely, web_traffic, Google_Index, and Links_pointing_to_page, were significantly associated with legitimate websites. These features remain robust against adversarial manipulation due to their dependency on external, distributed trust ecosystems (e.g., search engine authority and web visibility). This converges with the trust-based modelling paradigm advocated by van Geest et al. (2023) and positions metadata as a cornerstone for resilient phishing detection pipelines. Notably, the asymmetric interpretability of metadata, where presence affirms legitimacy, but absence does not equate to phishing, demands nuanced implementation in production environments.

The resulting logistic regression model is statistically rigorous, with a McFadden's R^2 of 67%. What distinguishes this contribution is its performance and model interpretability. In contrast to the prevailing reliance on opaque, black-box models, this work offers a transparent statistical framework that yields actionable insights, a critical requirement in cybersecurity domains where auditability, explainability, and risk governance are paramount. By leveraging odds ratios and feature significance, the model exposes the underlying anatomy of phishing strategies, providing both a detection mechanism and a knowledge base for cyber defense professionals.

Looking forward, this study opens multiple promising directions. Integrating this interpretable model with real-time systems, embedding SHAP-based feature attribution, and adapting to temporal changes in attacker behavior will further elevate its strategic value. Moreover, its transparent architecture makes it an ideal candidate for hybrid deployment alongside machine and deep learning models, enabling the best of both interpretability and adaptability.

REFERENCES

Adebowale, M. A., Lwin, K. T., & Hossain, M. A. (2022.). *Intelligent Phishing Detection Scheme Using Deep Learning Algorithms*. https://docs.apwg.org/reports/apwg_trends_report_q2_2019.pdf

Al-Ahmadi, S., Alotaibi, A., & Alsaleh, O. (2022). PDGAN: Phishing Detection With Generative Adversarial Networks. *IEEE Access*, *10*, 42459–42468. https://doi.org/10.1109/ACCESS.2022.3168235

Alazaidah, R., Al-Shaikh, A., AL-Mousa, M. R., Khafajah, H., Samara, G., Alzyoud, M., Al-Shanableh, N., & Almatarneh, S. (2024). Website Phishing Detection Using Machine Learning Techniques. *Journal of Statistics Applications and Probability*, *13*(1), 119–129. https://doi.org/10.18576/jsap/130108

Aljofey, A., Jiang, Q., Qu, Q., Huang, M., & Niyigena, J. P. (2020). An effective phishing detection model based on character level convolutional neural network from URL. *Electronics (Switzerland)*, *9*(9), 1–24. https://doi.org/10.3390/electronics9091514

Aljofey, A., Jiang, Q., Rasool, A., Chen, H., Liu, W., Qu, Q., & Wang, Y. (2022). An effective detection approach for phishing websites using URL and HTML features. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-10841-5

Alshingiti, Z., Alaqel, R., Al-Muhtadi, J., Haq, Q. E. U., Saleem, K., & Faheem, M. H. (2023). A Deep Learning-Based Phishing Detection System Using CNN, LSTM, and LSTM-CNN. *Electronics (Switzerland)*, *12*(1). https://doi.org/10.3390/electronics12010232

Atlam, H. F., & Oluwatimilehin, O. (2023). Business Email Compromise Phishing Detection Based on Machine Learning: A Systematic Literature Review. In *Electronics (Switzerland)* (Vol. 12, Issue 1). MDPI. https://doi.org/10.3390/electronics12010042

Bax, S., McGill, T., & Hobbs, V. (2021). Maladaptive behaviour in response to email phishing threats: The roles of rewards and response costs. *Computers and Security*, *106*. https://doi.org/10.1016/j.cose.2021.102278

Cahusac, P. (2022). *Log Likelihood Ratios for Common Statistical Tests Using the likelihoodR Package*.

Catal, C., Giray, G., Tekinerdogan, B., Kumar, S., & Shukla, S. (2022). Applications of deep learning for phishing detection: a systematic literature review. *Knowledge and Information Systems*, *64*(6), 1457–1500. https://doi.org/10.1007/s10115-022-01672-x

Gandotra, E., & Gupta, D. (2021). *An Efficient Approach for Phishing Detection using Machine Learning* (pp. 239–253). https://doi.org/10.1007/978-981-15-8711-5_12

Hassan, S., Ahmad, R., Katuk, N., Ghazali, N. N., Aripin, J. A., & Ali, F. (2024). Staying One Step Ahead: Exploring Protection Motivation Theory to Combat Cyber-fraud Among E-services Users. *Procedia Computer Science*, *234*, 1364–1371. https://doi.org/10.1016/j.procs.2024.04.011

Mughaid, A., AlZu'bi, S., Hnaif, A., Taamneh, S., Alnajjar, A., & Elsoud, E. A. (2022). An intelligent cyber security phishing detection system using deep learning techniques. *Cluster Computing*, *25*(6), 3819–3828. https://doi.org/10.1007/s10586-022-03604-4

Nattino, G., Pennell, M. L., & Lemeshow, S. (2020). Assessing the goodness of fit of logistic regression models in large samples: A modification of the Hosmer-Lemeshow test. *Biometrics*, *76*(2), 549–560. https://doi.org/10.1111/biom.13249

Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications*, *117*, 345–357. https://doi.org/10.1016/j.eswa.2018.09.029

Sahingoz, O. K., Buber, E., & Kugu, E. (2024). DEPHIDES: Deep Learning Based Phishing Detection System. *IEEE Access*, *12*, 8052–8070. https://doi.org/10.1109/ACCESS.2024.3352629

Senaviratna, N. A. M. R., & Cooray, T. M. J. A. (2019). *Detecting Multicollinearity of Binary Logistic Regression Model: An Analysis of Motorcycle Accidents in Sri Lanka*.

Shahrivari, V., Darabi, M. M., & Izadi, M. (2020). *Phishing Detection Using Machine Learning Techniques*. http://arxiv.org/abs/2009.11116

Silva, C. M. R. da, Feitosa, E. L., & Garcia, V. C. (2020). Heuristic-based strategy for Phishing prediction: A survey of URL-based approach. *Computers and Security*, *88*. https://doi.org/10.1016/j.cose.2019.101613

Sonowal, G., & Kuppusamy, K. S. (2020). PhiDMA – A phishing detection model with multi-filter approach. *Journal of King Saud University - Computer and Information Sciences*, *32*(1), 99–112. https://doi.org/10.1016/j.jksuci.2017.07.005

Tan, C. C. L., Chiew, K. L., Yong, K. S. C., Sebastian, Y., Than, J. C. M., & Tiong, W. K. (2023). Hybrid phishing detection using joint visual and textual identity. *Expert Systems with Applications*, *220*. https://doi.org/10.1016/j.eswa.2023.119723

Toptancı, Ş., Erginel, N., & Acar, I. (2023). Predicting the severity of occupational accidents in the construction industry using standard and regularized logistic regression models İnşaat sektöründe standart ve düzenlileştirilmiş lojistik regresyon modelleri kullanılarak iş kazalarının şiddetinin tahmini. *Bilim. Derg. / NOHU J. Eng. Sci*, *12*(3), 778–798. https://doi.org/10.28948/ngmuh.1212385

Ugba, E. R., & Gertheiss, J. (2023). A modification of McFadden's R2 for binary and ordinal response models. *Communications for Statistical Applications and Methods*, *30*(1), 49–63. https://doi.org/10.29220/CSAM.2023.30.1.049

Van Dooremaal, B., Burda, P., Allodi, L., & Zannone, N. (2021, August 17). Combining Text and Visual Features to Improve the Identification of Cloned Webpages for Early Phishing Detection. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3465481.3470112

van Geest, R. J., Cascavilla, G., Hulstijn, J., & Zannone, N. (2024). The applicability of a hybrid framework for automated phishing detection. *Computers and Security*, *139*. https://doi.org/10.1016/j.cose.2024.103736

Vijayalakshmi, M., Mercy Shalinie, S., Yang, M. H., & Raja Meenakshi, U. (2020). Web phishing detection techniques: A survey on the state-of-the-art, taxonomy and future directions. In *IET Networks* (Vol. 9, Issue 5, pp. 235–246). Institution of Engineering and Technology. https://doi.org/10.1049/iet-net.2020.0078

Wei, W., Ke, Q., Nowak, J., Korytkowski, M., Scherer, R., & Woźniak, M. (2020). Accurate and fast URL phishing detector: A convolutional neural network approach. *Computer Networks*, *178*. https://doi.org/10.1016/j.comnet.2020.107275

Yang, P., Zhao, G., & Zeng, P. (2019). Phishing website detection based on multidimensional features driven by deep learning. *IEEE Access*, *7*, 15196–15209. https://doi.org/10.1109/ACCESS.2019.2892066

Yao, W., Ding, Y., & Li, X. (2018). Deep Learning for Phishing Detection. *2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/Sustain Com)*, 645–650. https://doi.org/10.1109/BDCloud.2018.00099

Yavartanoo, F., Brossard, M., Bull, S. B., Paterson, A. D., & Yoo, Y. J. (2025). Dimension Reduction Using Local Principal Components for Regression-Based Multi-SNP Analysis in 1000 Genomes and the Canadian Longitudinal Study on Aging (CLSA). *Genetic Epidemiology*, *49*(3). https://doi.org/10.1002/gepi.70005